Análisis Inteligente de Datos: Segundo Parcial

Claudio Sebastián Castillo

10 de mayo de 2022

ANOVA

Datos

Observaciones por grupo:

Se cumplen los supuestos para su implementación?

Anova

fit del modelo

coeficientes

p-value

F-value

Plot ANOVA

Conclusión

Testear homosedasticidad

Test de Bartlett

sensibilidad al supuesto de normalidad

Testear normalidad

Testear normalidad analizando residuos

Anova y después: post-hoc

Tukey's Honest Significant Differences (HSD)

Cuando ANOVA no funciona: test de Kruskal-Wallis

ANOVA_multivariante

Datos

Gafico

Test Anovam

Tamaño del efecto

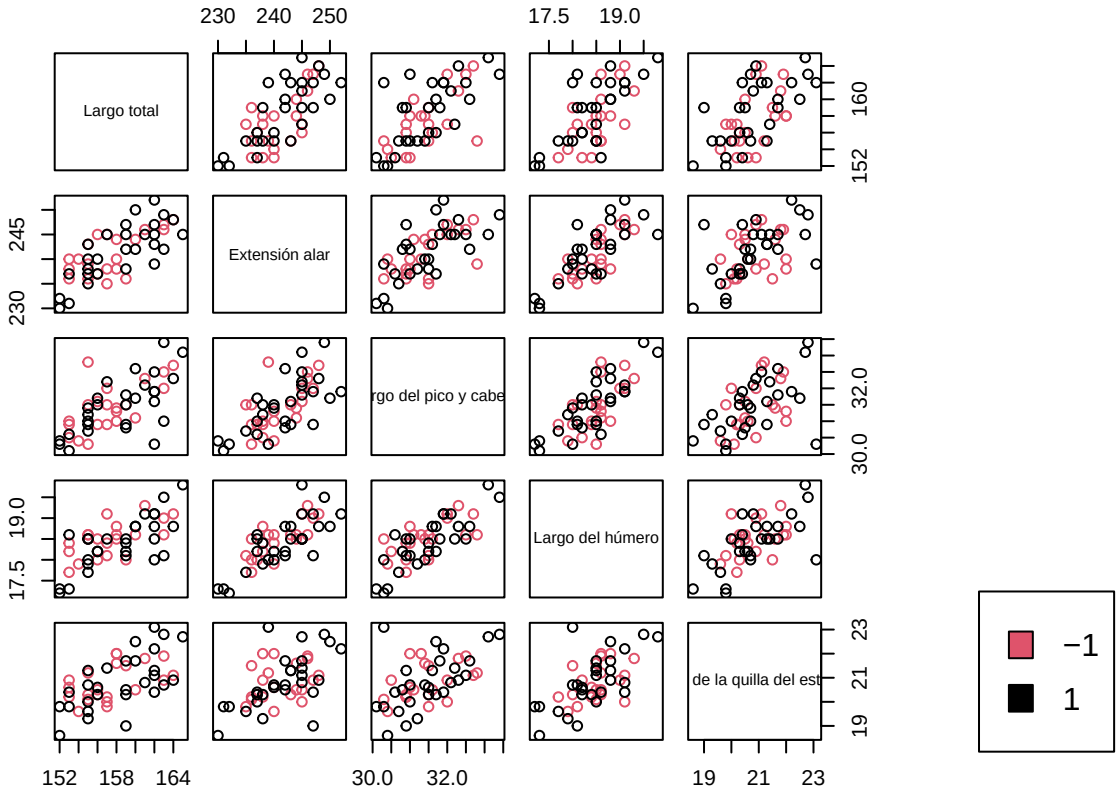
.01: Small effect size .06: Medium effect size .14 or higher: Large effect size

Analisis Discriminante Lineal (LDA)

Datos

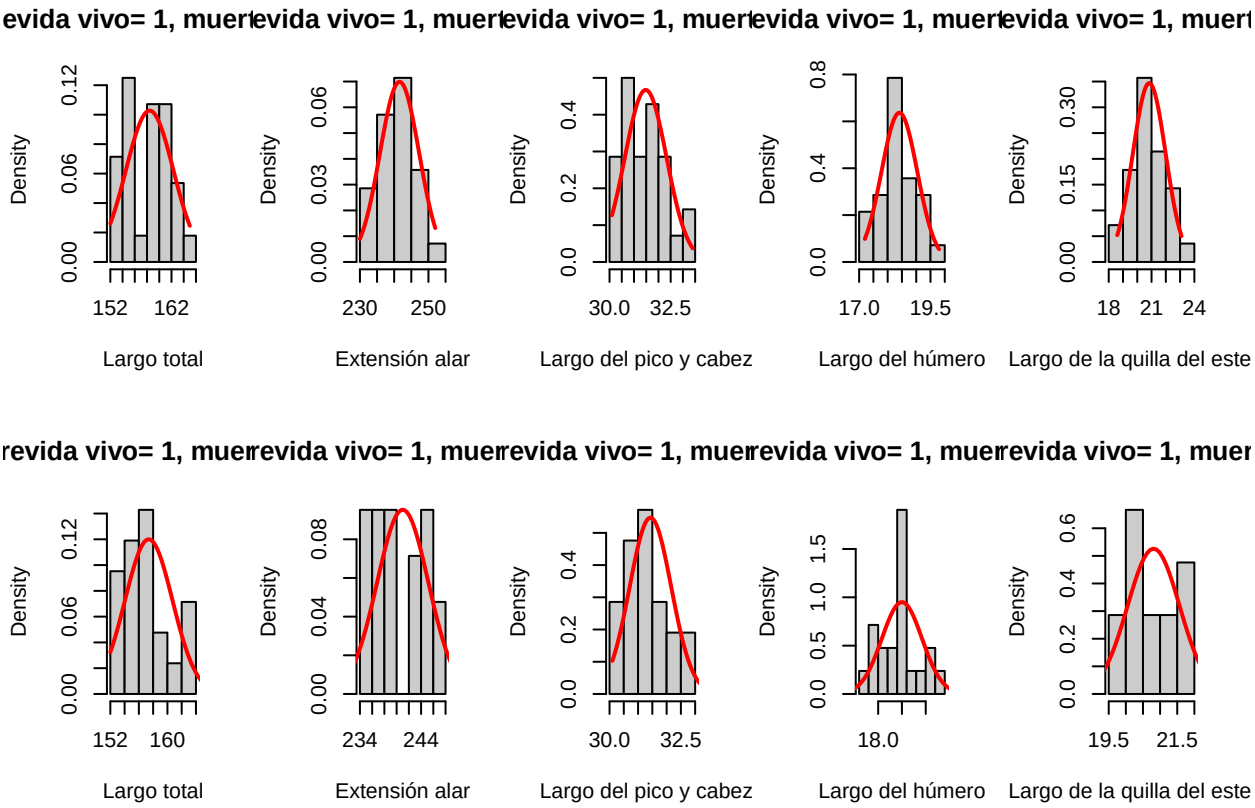
```
## tibble [49 x 6] (S3: tbl_df/tbl/data.frame)
##  $ Largo total                    : num [1:49] 156 154 153 153 155 163 157 155 164 158 ...
##  $ Extensión alar                 : num [1:49] 245 240 240 236 243 247 238 239 248 238 ...
##  $ Largo del pico y cabeza        : num [1:49] 31.6 30.4 31 30.9 31.5 32 30.9 32.8 32.7 31 ...
##  $ Largo del húmero               : num [1:49] 18.5 17.9 18.4 17.7 18.6 19 18.4 18.6 19.1 18.8 ...
##  $ Largo de la quilla del esternón: num [1:49] 20.5 19.6 20.6 20.2 20.3 20.9 20.2 21.2 21.1 22 ...
##  $ Sobrevida vivo= 1, muerto= -1  : Factor w/ 2 levels "-1","1": 2 2 2 2 2 2 2 2 2 2 ...
```

Explorando discriminación por pares de variable



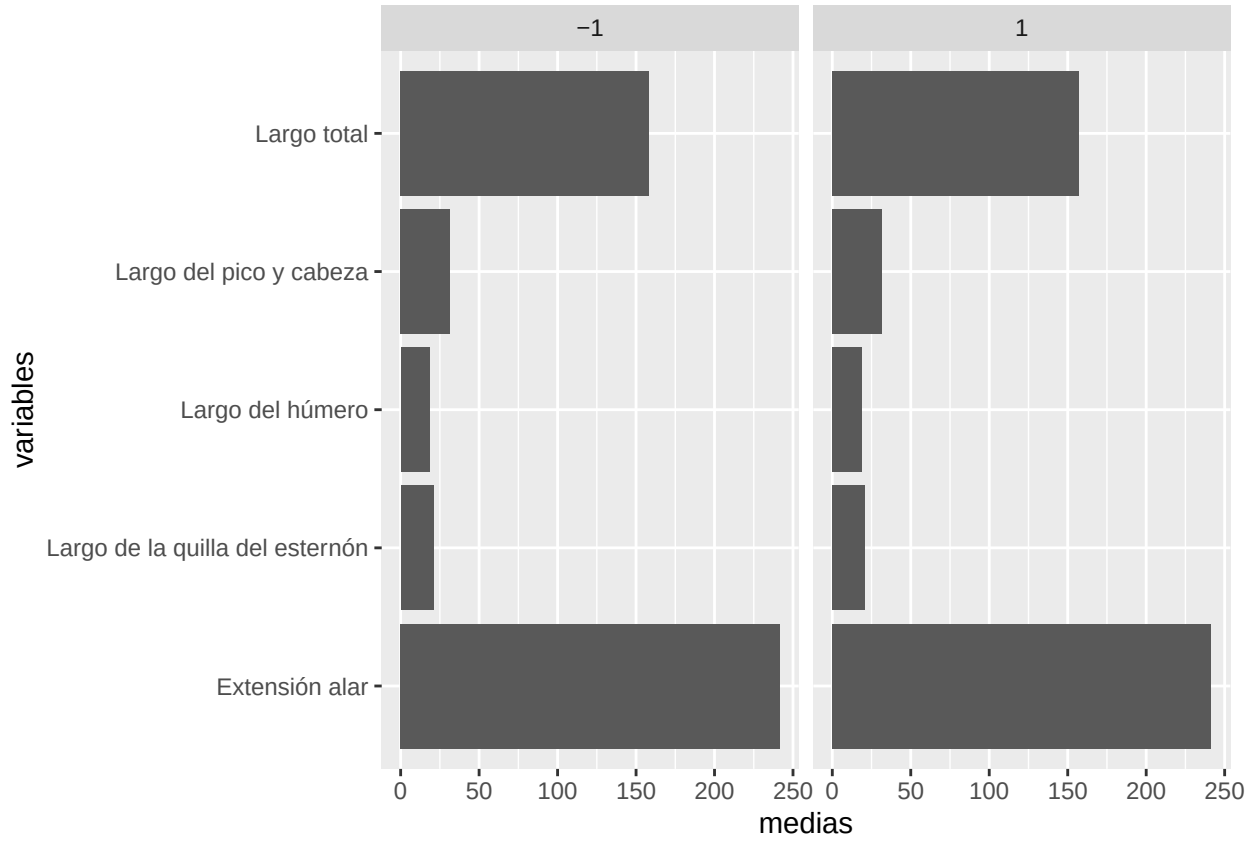
Ver qué par de variables separa bien las Sobrevida vivo= 1, muerto= -1

Homogeneidad de la Varianza: Histograma VariablexGrupo



Medias de las variables por Grupo

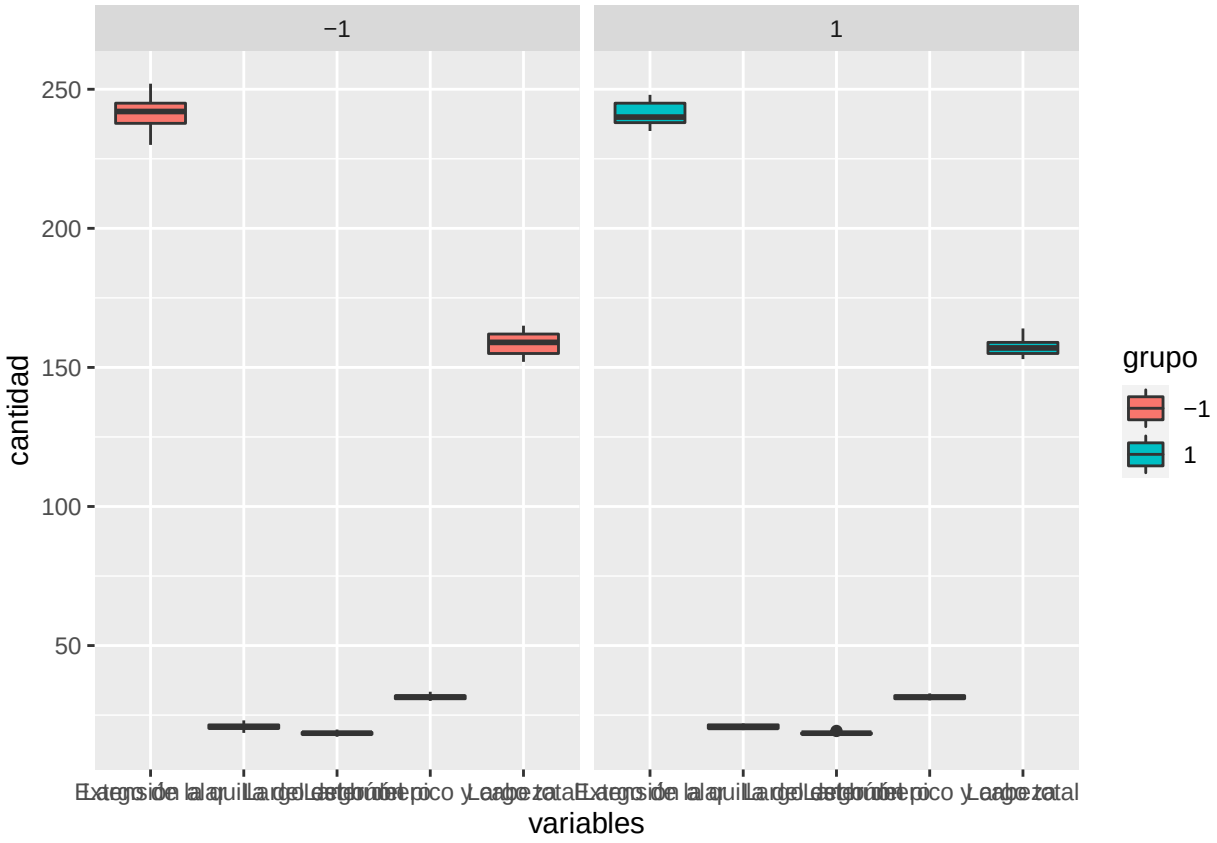
Las variables son muy parecidas.


−1
1
Largo total
Largo del pico y cabeza
Largo del húmero
Largo de la quilla del esternón
Extensión alar
variables
0
50
100
150
200
250
medias

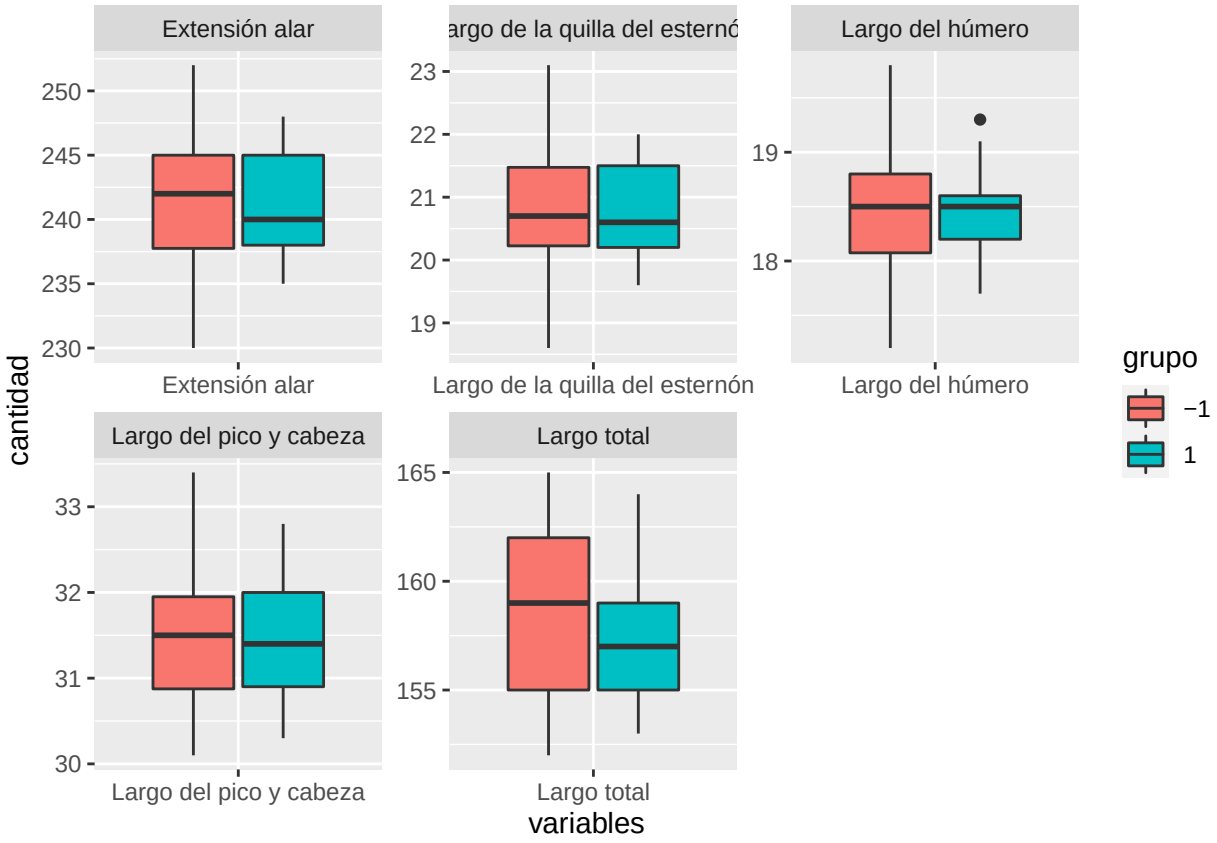
Boxplot

Boxplot todas las variables

Box por grupo



Box por variable



Vectores medios de ambos grupos

```
## # A tibble: 10 x 3
## # Groups:   grupo [2]
##    grupo variables                      medias
##    <fct> <chr>                           <dbl>
##  1 -1    Extensión alar                  242.
##  2 -1    Largo de la quilla del esternón  20.8
##  3 -1    Largo del húmero                 18.4
##  4 -1    Largo del pico y cabeza          31.5
##  5 -1    Largo total                     158.
##  6 1     Extensión alar                  241
##  7 1     Largo de la quilla del esternón  20.8
##  8 1     Largo del húmero                 18.5
##  9 1     Largo del pico y cabeza          31.4
## 10 1     Largo total                     157.
```

Viendo los gráficos de las distribuciones de datos en las variables por grupo y los respectivos vectores medios no parece útil realizar un análisis discriminante.

Subseleccionamos variables para el análisis

Seleccionamos para el análisis las variables cuyas medias resultan más discriminantes: largo total y extension alar

Contraste de Normalidad Univariante Shapiro-Wilk

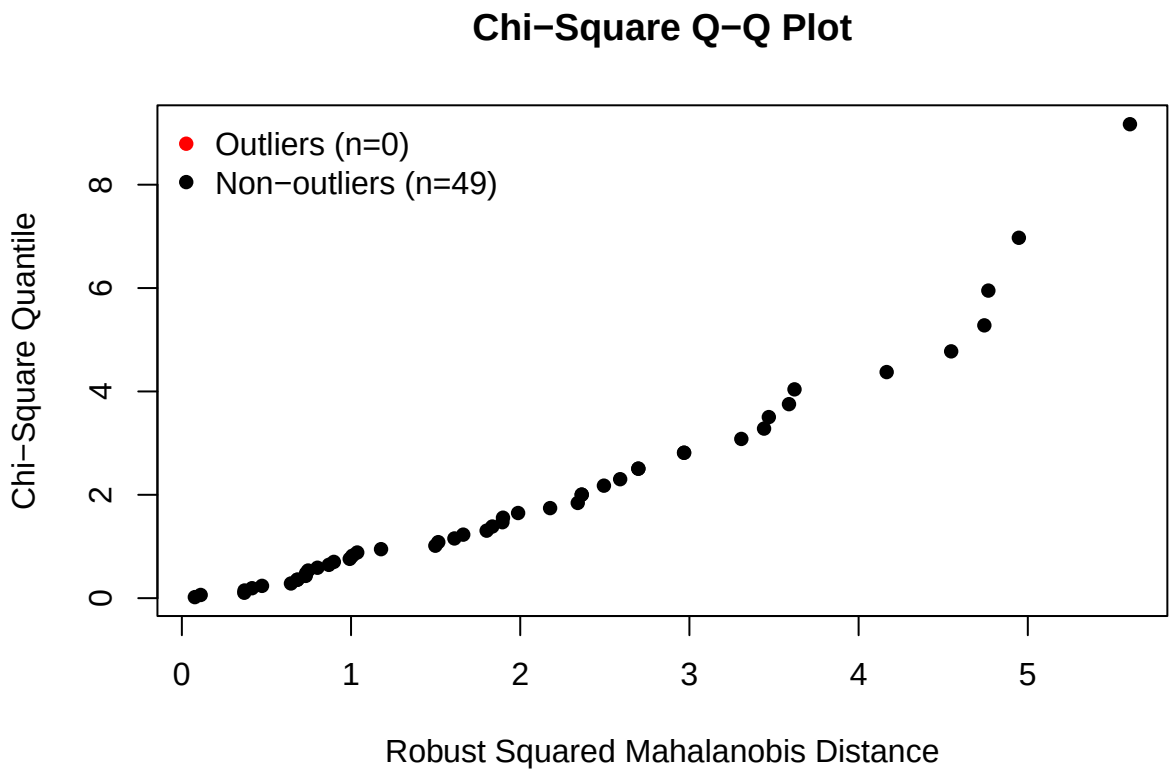
datos_tidy[["Sobrevida vivo= 1, muerto= -1"]]	variable	p_value_Shapiro.test
-1	Largo total	0.10038
-1	Extensión alar	0.70606
1	Largo total	0.16530
1	Extensión alar	0.08237

```
## [1] "No hay evidencia de falta de normalidad univariante en ninguna variable predictora por grupo"
```

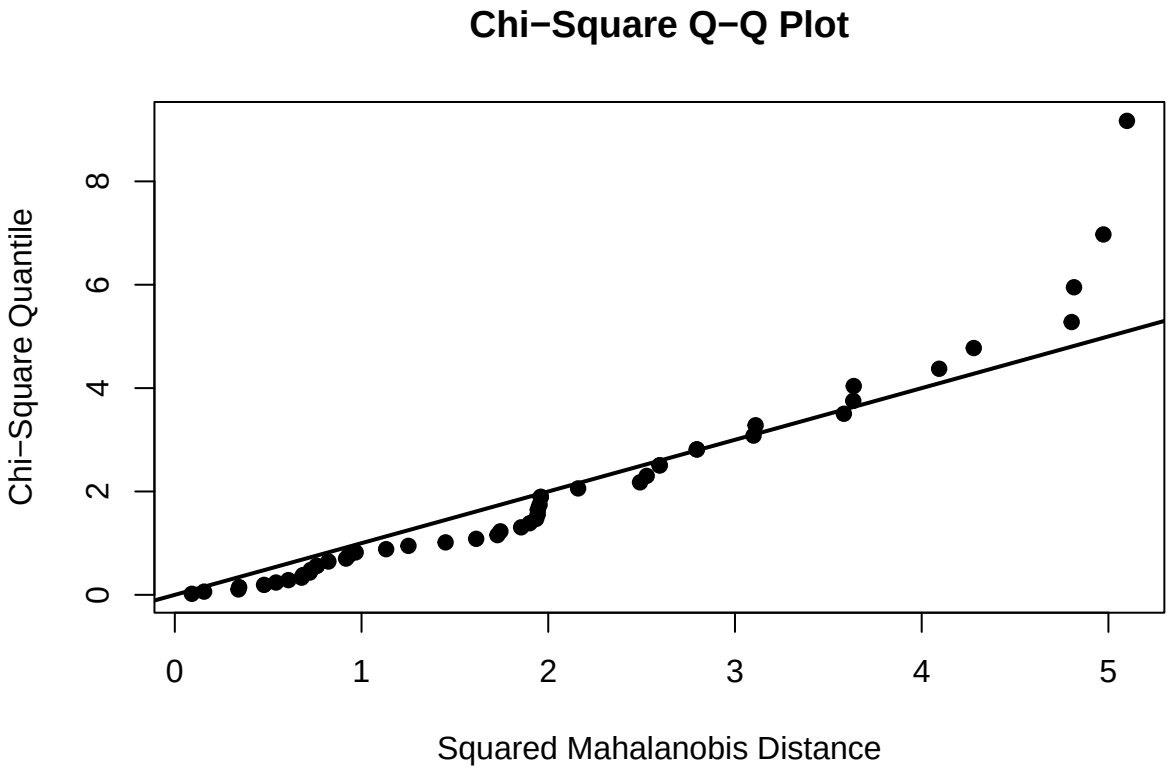
```
## # A tibble: 0 x 3
## # Groups:   datos_tidy[["Sobrevida vivo= 1, muerto= -1"]] [0]
## # ... with 3 variables: datos_tidy[["Sobrevida vivo= 1, muerto= -1"]] <fct>,
## #   variable <fct>, p_value_Shapiro.test <dbl>
```

Contraste de Normalidad MultiVariante

Outliers



Test de Royston




```
##      Test        H   p value MVN
## 1 Royston 4.457852 0.1009865 YES
```

```
## [1] "No hay evidencia de falta de normalidad multivariante a nivel de significancia 0.05"
```

Test de Henze-Zirkler

```
##            Test        HZ   p value MVN
## 1 Henze-Zirkler 0.6637743 0.1823398 YES
```

```
## [1] "No hay evidencia de falta de normalidad multivariante a nivel de significancia 0.05 "
```

Contraste de Matriz de Covarianza

```
##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  temp
## Chi-Sq (approx.) = 2.0318, df = 3, p-value = 0.5658
```

```
## [1] "Se puede aceptar que la matriz de covarianza es igual en todos los grupos"
```

Estimación de parámetros de la función de densidad (uˆ(X),E) y cálculo de la función discriminante según aproximación de Fisher via lda()

```
## Call:
```

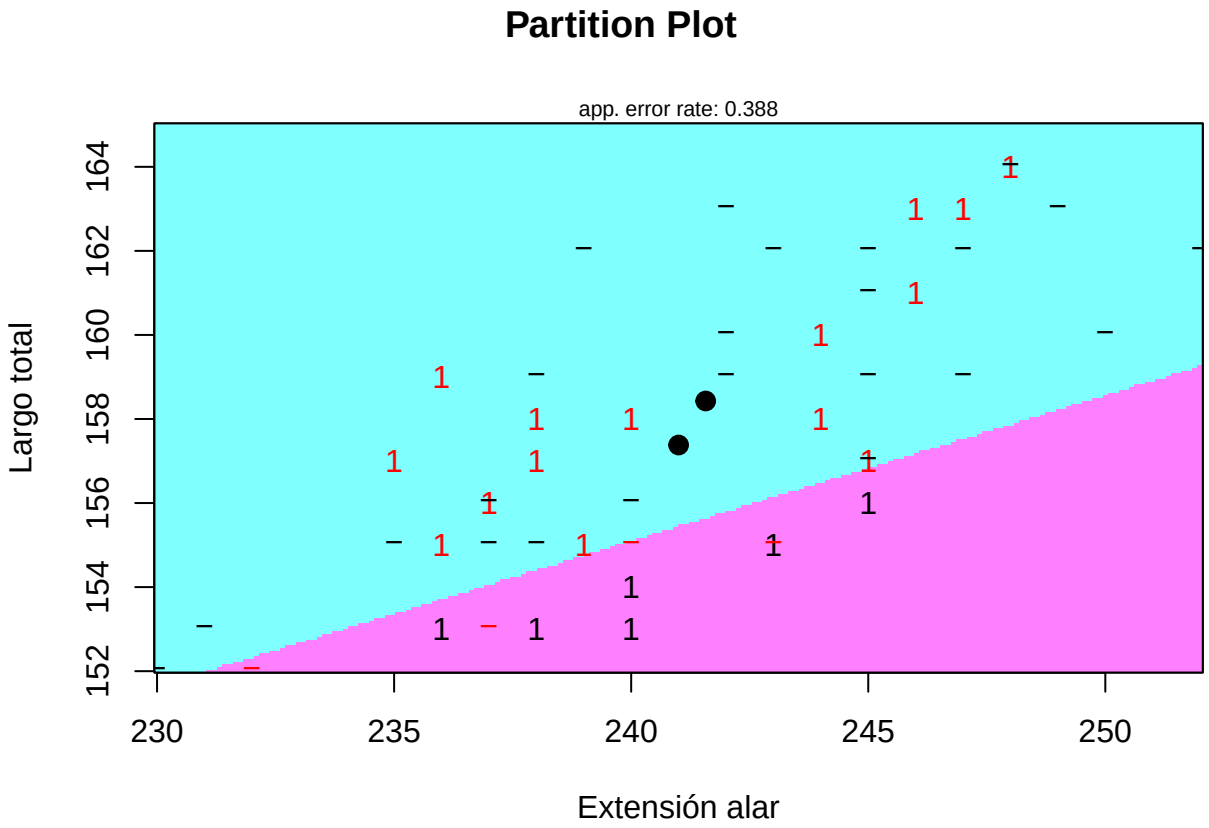
```
## lda(temp, datos[[{
##     {
##         variable_factor_lda
##     }
## }]])
##
## Prior probabilities of groups:
##        -1         1
## 0.5714286 0.4285714
##
## Group means:
##    Largo total Extensión alar
## -1    158.4286       241.5714
## 1     157.3810       241.0000
##
## Coefficients of linear discriminants:
##                       LD1
## Largo total    -0.3787943
## Extensión alar  0.1312623
```

Evaluación del error: Accuracy Table

```
##           Clase predicha
## Clase real -1  1
##         -1 25  3
##         1  15  6
```

```
## [1] "trainig_error = 36.734693877551 %"
```

Visualización de las clasificaciones



Analisis Discriminante Cuadrático (QDA)>falta de homocedasticidad/outliers LDA

Explorando discriminación por pares de variable

Contraste de Normalidad Univariante Shapiro-Wilk

Contraste de Normalidad MultiVariante

Outliers

Test de Royston

Test de Henze-Zirkler

Contraste de Matriz de Covarianza

Parámetros de la función de densidad función discriminante según aproximación de Fisher via qda()

Evaluación del error: Accuracy Table

Visualización de las clasificaciones

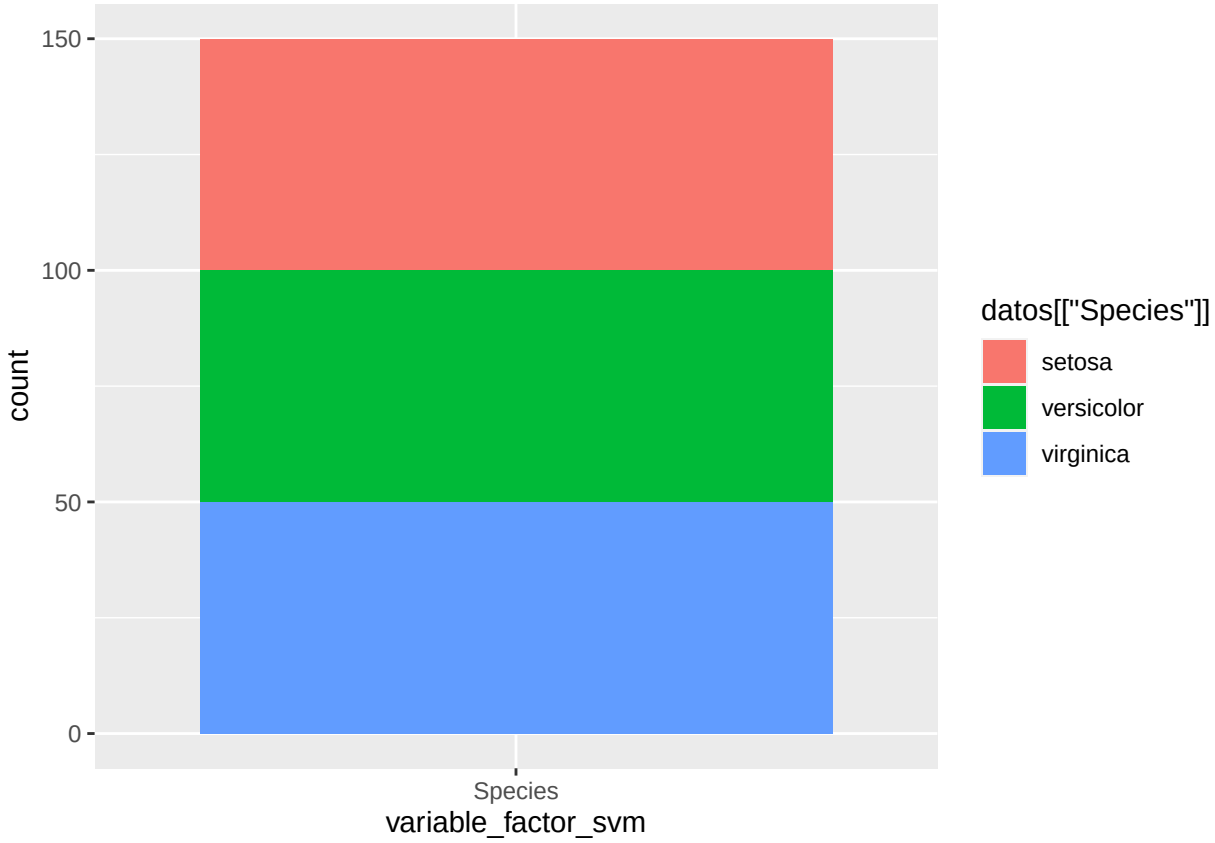
Analisis Discriminante Cuadrático Robusto (RQDA)>falta normalidad

Máquinas de Soporte Vectorial

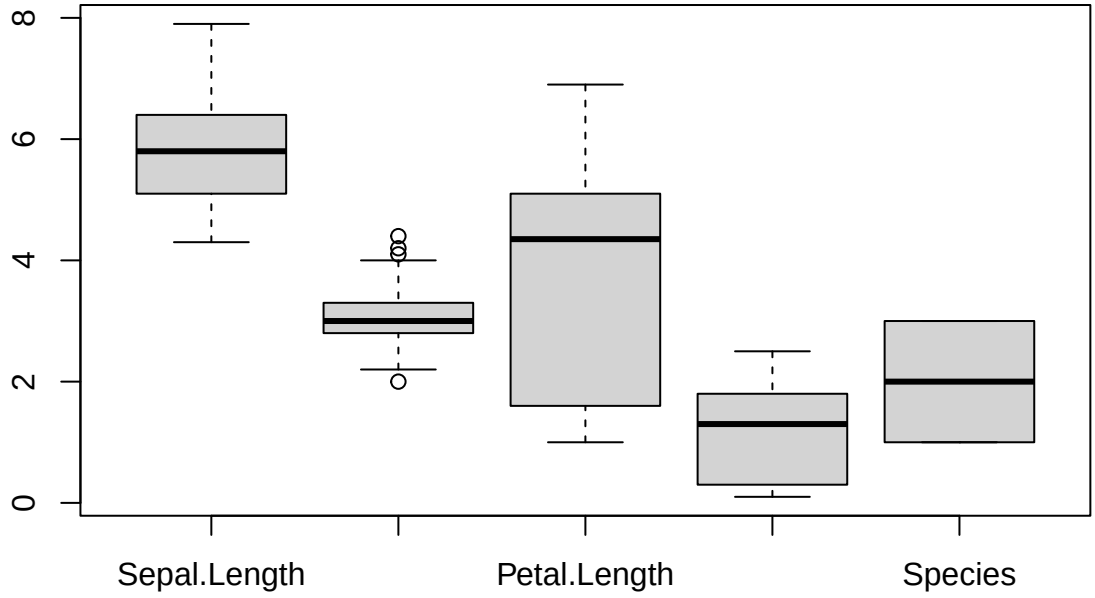
Datos

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

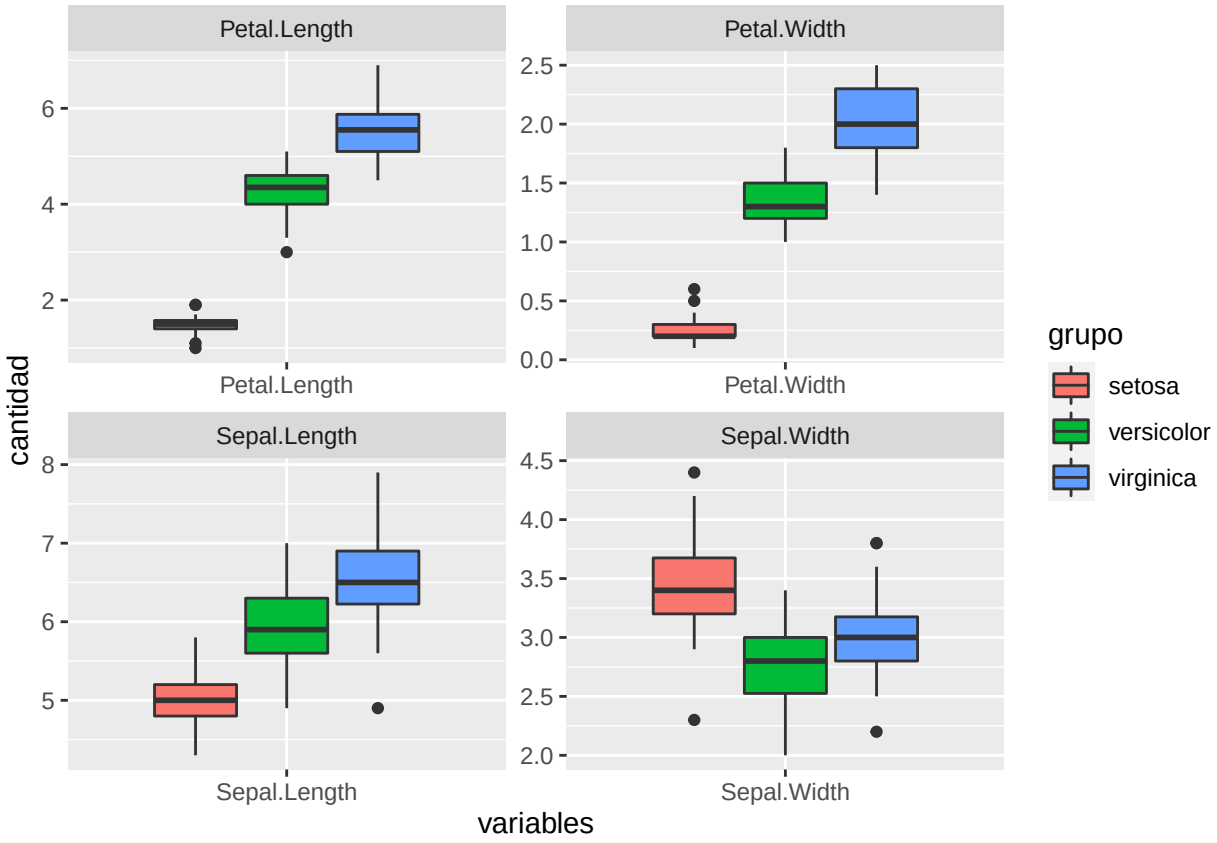
Grafico datos



Boxplot



Box por variable



(a) Analice cuales valores medios son diferentes en las especies: petal length and width

```
## [1] 120   5
```

```
## [1] 30  5
```

Busqueda de mejor hiperparametro C (coste)

```
##
## Parameter tuning of 'svm':
##
## - sampling method: 10-fold cross validation
##
## - best parameters:
##  cost
##     1
##
## - best performance: 0.01666667
##
## - Detailed performance results:
##     cost      error dispersion
## 1  0.001 0.81666667 0.07657805
## 2  0.010 0.13333333 0.09781565
## 3  0.100 0.04166667 0.04392052
## 4  1.000 0.01666667 0.03513642
## 5  5.000 0.04166667 0.04392052
```

```
## 6 10.000 0.03333333 0.04303315
## 7 15.000 0.05000000 0.04303315
## 8 20.000 0.05000000 0.04303315
```

Mejor modelo según hiperparametro

```
##
## Call:
## best.tune(method = svm, train.x = temp, train.y = datos_train[[{
##     {
##         variable_factor_svm
##     }
## }]], ranges = list(cost = c(0.001, 0.01, 0.1, 1, 5, 10, 15, 20)),
##     kernel = "linear", scale = TRUE)
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  linear
##        cost:  1
##
## Number of Support Vectors:  25
##
##  ( 2 13 10 )
##
##
## Number of Classes:  3
##
## Levels:
##  setosa versicolor virginica
```

```
## [1] 19 34 42 44 46 47
```

Predicciones del Modelo

```
##             real
## prediccion   setosa versicolor virginica
##   setosa         10          0         0
##   versicolor      0         10         1
##   virginica       0          0         9
```

```
## [1] "Observaciones de test mal clasificadas: 3.33 %"
```

```
## [1] "Observaciones de test bien clasificadas: 96.67 %"
```

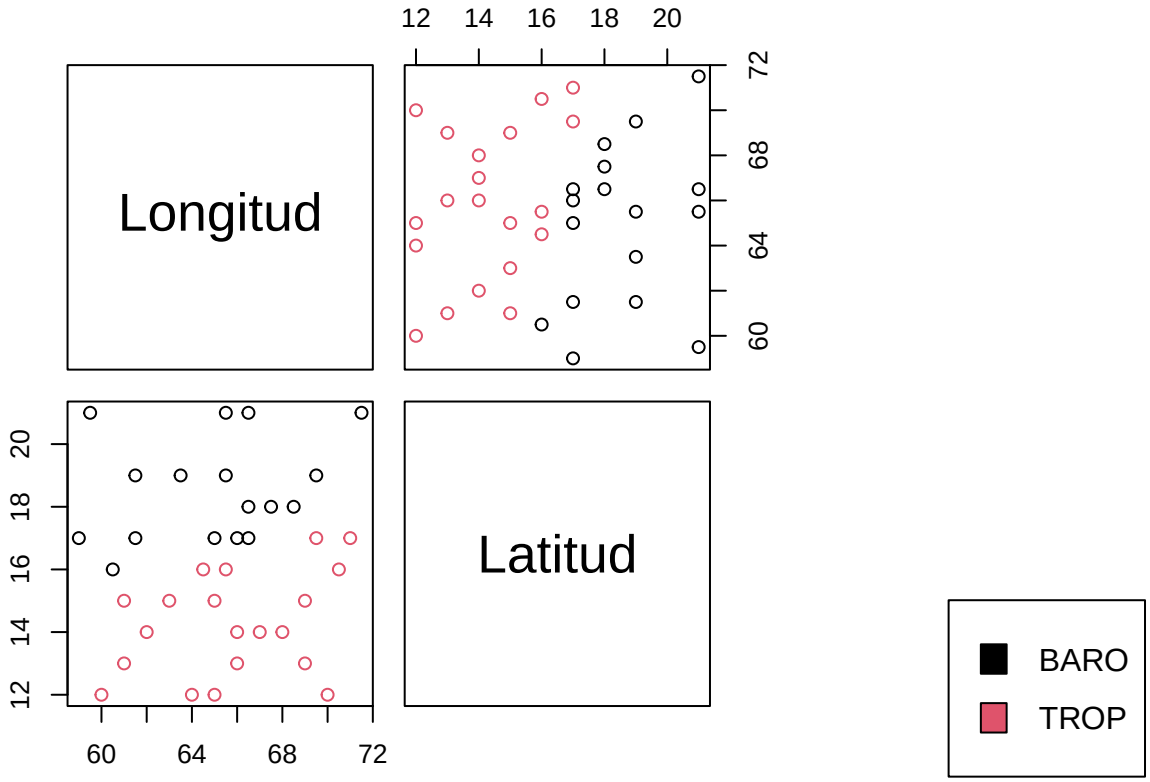
Los bien clasificados por especies son: 100% en setosa y virginica, y 90% en versicolor. El error en versicolor y virgínica es lógico porque son los grupos con mayor similitud en sus datos.

Analisis Discriminante Lineal (LDA)

Datos

```
## tibble [37 x 3] (S3: tbl_df/tbl/data.frame)
##  $ Longitud: num [1:37] 59 59.5 60 60.5 61 61 61.5 61.5 62 63 ...
##  $ Latitud : num [1:37] 17 21 12 16 13 15 17 19 14 15 ...
##  $ Clase   : Factor w/ 2 levels "BARO","TROP": 1 1 2 1 2 2 1 1 2 2 ...
```

Explorando discriminación por pares de variable

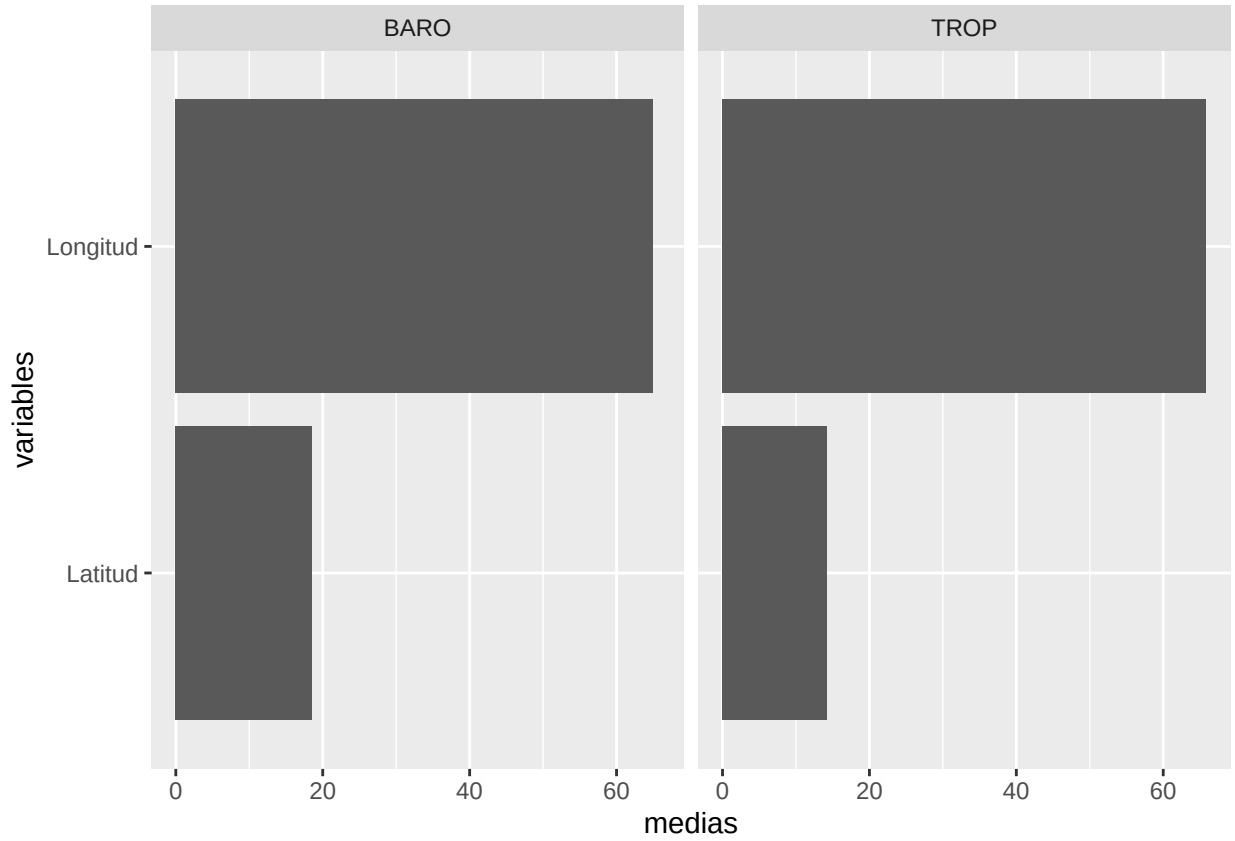


Ver qué par de variables separa bien las Clase

Homogeneidad de la Varianza: Histograma VariablexGrupo

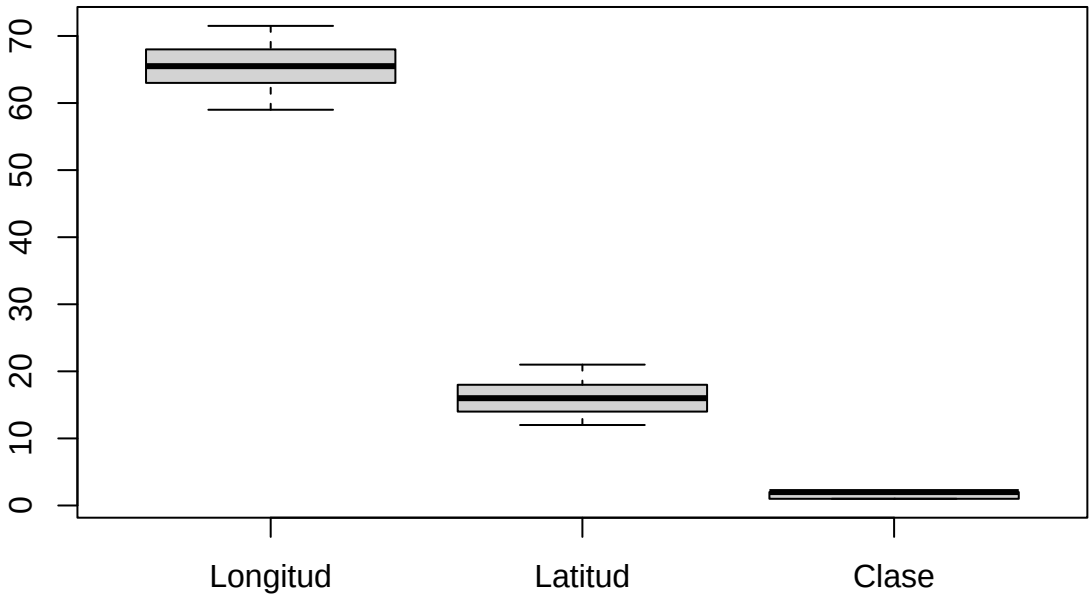
Medias de las variables por Grupo

Las variables son muy parecidas.

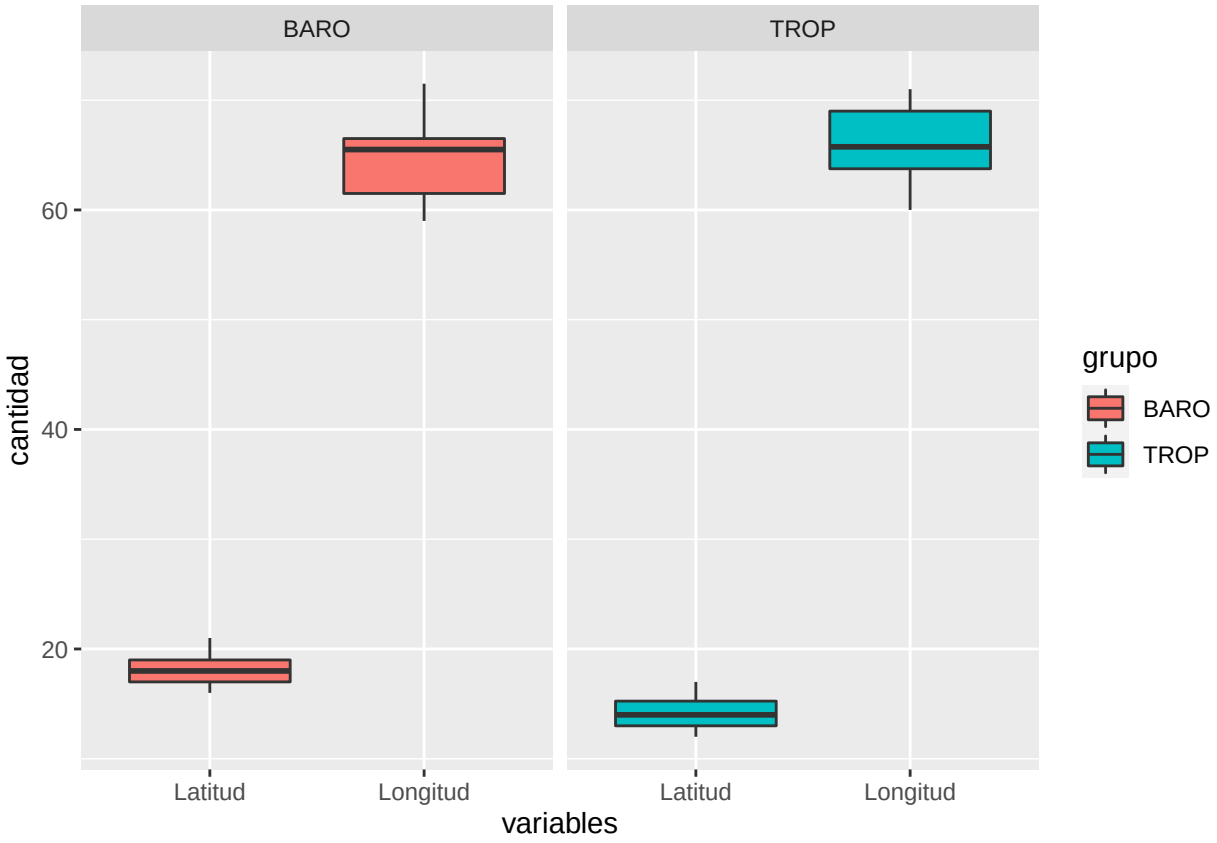


Boxplot

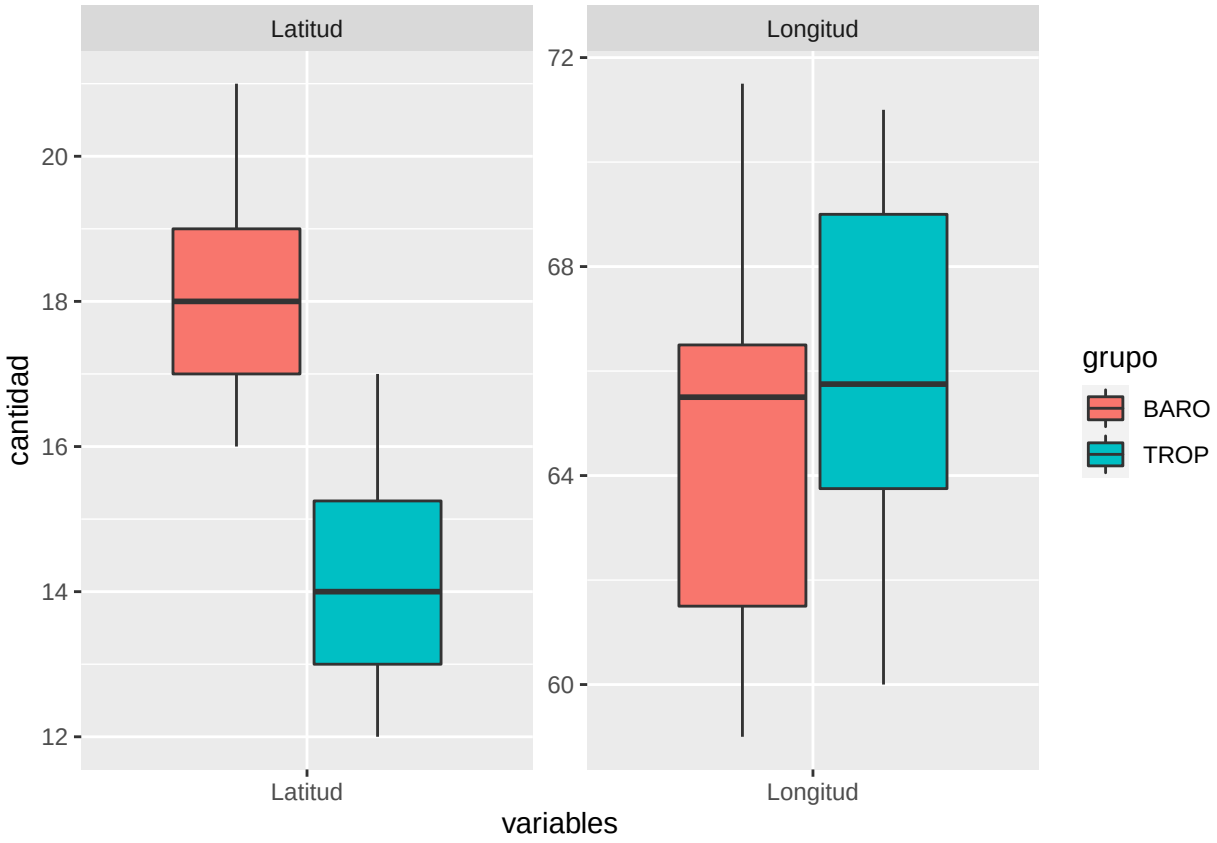
Boxplot todas las variables



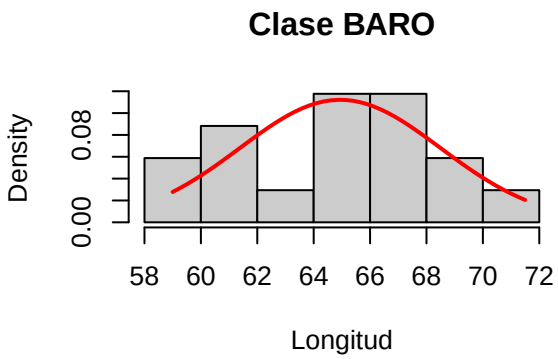
Box por grupo

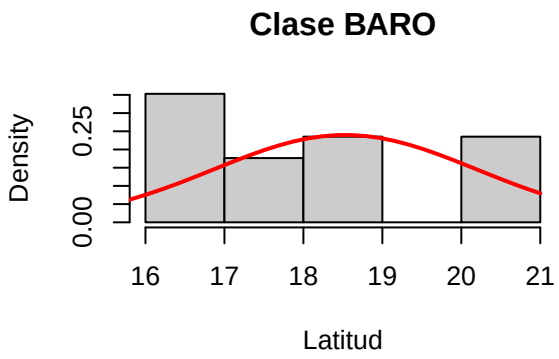


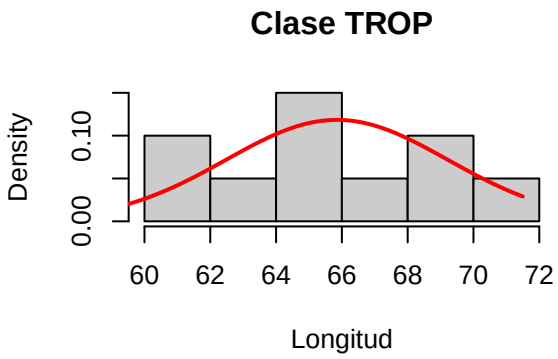
Box por variable

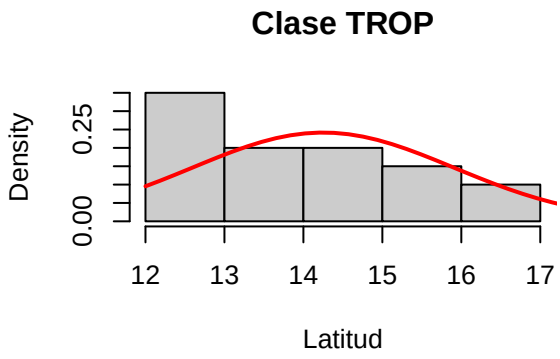


Histogramas









Vectores medios de ambos grupos

```
## # A tibble: 4 x 3
## # Groups:   grupo [2]
##   grupo variables medias
##   <fct> <chr>      <dbl>
## 1 BARO  Latitud     18.5
## 2 BARO  Longitud    64.9
## 3 TROP  Latitud     14.2
## 4 TROP  Longitud    65.8
```

Contraste de Normalidad Univariante Shapiro-Wilk

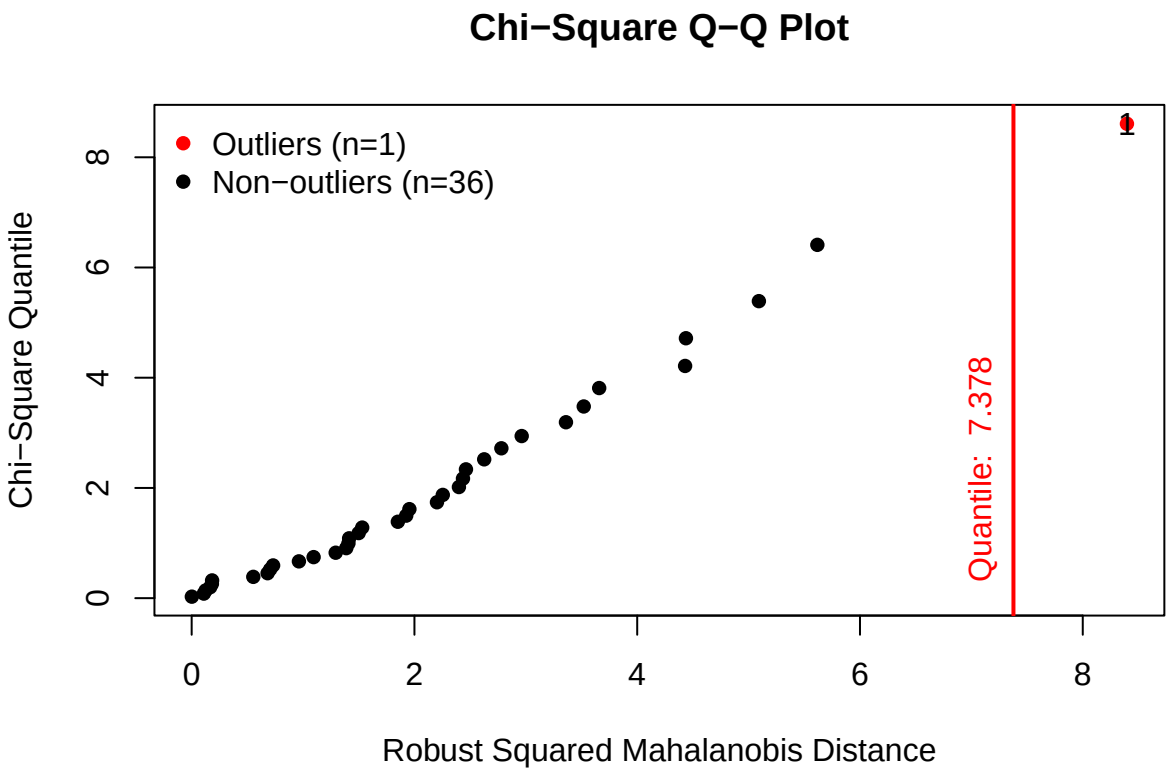
datos_tidy[["Clase"]]	variable	p_value_Shapiro.test
BARO	Longitud	0.59928
BARO	Latitud	0.03158
TROP	Longitud	0.46410
TROP	Latitud	0.12715

```
## [1] "H0 debe rechazarse: hay evidencia de falta de normalidad en los siguientes casos"
```

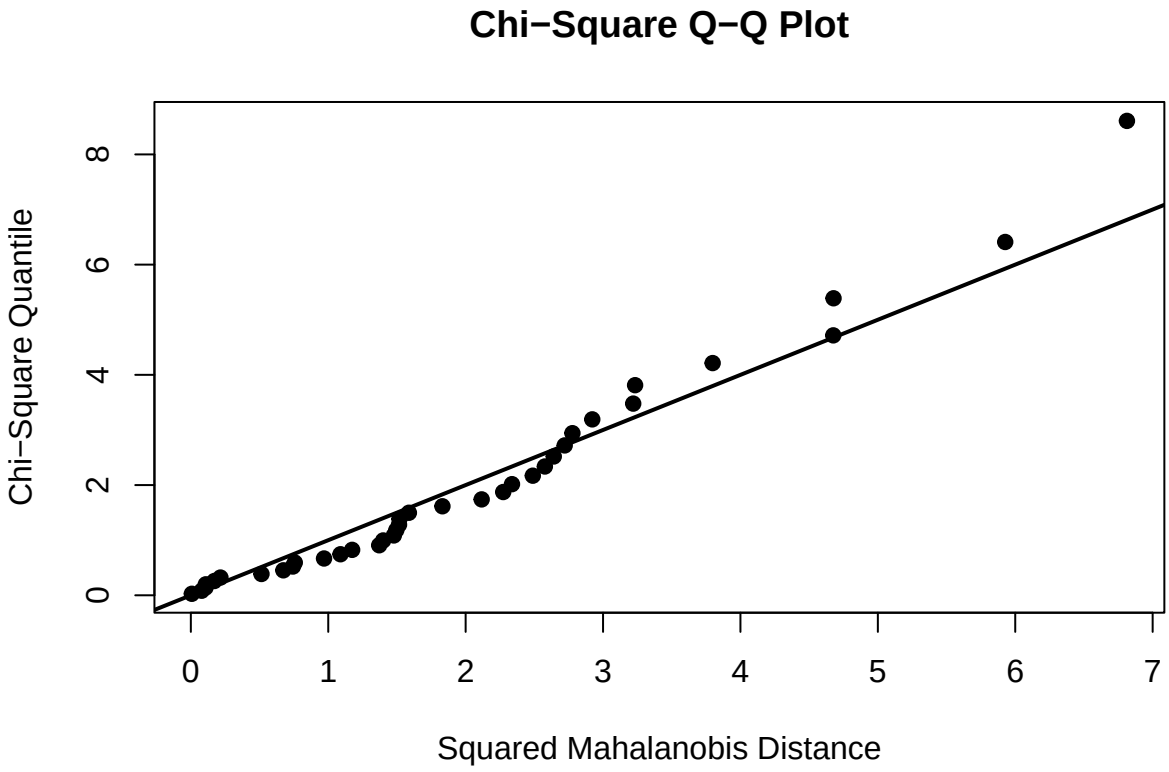
```
## # A tibble: 1 x 3
## # Groups:   datos_tidy[["Clase"]] [1]
##   `datos_tidy[["Clase"]]` variable p_value_Shapiro.test
##   <fct>                   <fct>                   <dbl>
## 1 BARO                    Latitud                0.0316
```

Contraste de Normalidad MultiVariante

Outliers



Test de Royston




```
##      Test        H   p value MVN
## 1 Royston 3.764152 0.1522745 YES
```

```
## [1] "No hay evidencia de falta de normalidad multivariante a nivel de significancia 0.05"
```

Test de Henze-Zirkler

```
##            Test        HZ   p value MVN
## 1 Henze-Zirkler 0.1868302 0.9883121 YES
```

```
## [1] "No hay evidencia de falta de normalidad multivariante a nivel de significancia 0.05 "
```

Contraste de Matriz de Covarianza

```
##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  temp
## Chi-Sq (approx.) = 0.15631, df = 3, p-value = 0.9843
```

```
## [1] "Se puede aceptar que la matriz de covarianza es igual en todos los grupos"
```

Estimación de parámetros de la función de densidad (u^(X),E) y cálculo de la función discriminante según aproximación de Fisher via lda()

```
## Call:
```

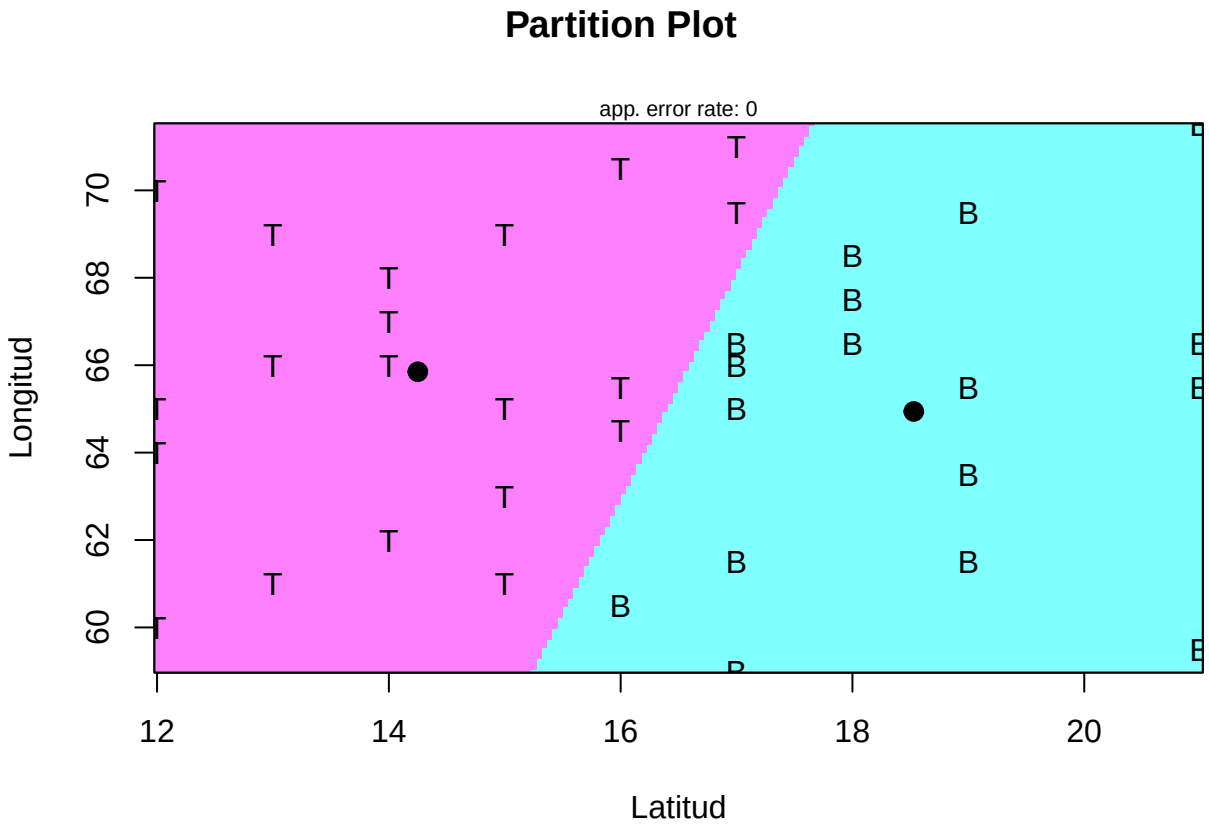
```
## lda(temp, datos[[{
##     {
##         variable_factor_lda
##     }
## }]])
##
## Prior probabilities of groups:
##      BARO      TROP
## 0.4594595 0.5405405
##
## Group means:
##      Longitud  Latitud
## BARO 64.94118 18.52941
## TROP 65.85000 14.25000
##
## Coefficients of linear discriminants:
##                 LD1
## Longitud  0.1220731
## Latitud  -0.6331236
```

Evaluación del error: Accuracy Table

```
##           Clase predicha
## Clase real BARO TROP
##       BARO   17    0
##       TROP    0   20
```

```
## [1] "trainig_error = 0 %"
```

Visualización de las clasificaciones



El test no parece significativo, las observaciones son linealmente separables. No hay normalidad univariante según el Test de Shapiro.